BY MAUREEN NEVIN

Computing Like Life Depends on It

TGen teams IBM System p and System x servers with Linux technology to facilitate cancer-gene research

It seems everyone has known someone who computes like her life—or someone else's—depends on it. But at The Translational Genomics Research Institute (TGen), in Phoenix, where the study of billions of possible gene interactions can be compressed from 12 months down to one week, the impetus really is life. The scientists at TGen work to find patterns and interactions that can signal cancer-causing genes, which could lead to cancer-killing drugs. The faster they can comb through the haystacks of gene combinations for the desired patterns, the greater the hope for the eventual vaccine or treatment that will save a patient's life.

In research situations, it's important to have powerful computing ability on machines that run well in an open-systems environment—sharing files and running parallel applications. Why open? Because researchers in this field are continuously writing new programs and upgrading old ones to quickly exploit their latest findings. A computer program to test a new theory that a type of cancer cell may express a particular gene or protein pattern could've been created on any number of machines. Rather than waste time and risk the possibility of conversion errors, open environments are the practical answer. That's why most researchers, who are usually fighting budget as well time constraints, write in C on an open-source system.

Open systems in use at the ASU-TGen High Performance Computing Center, at Arizona State University (ASU), includes an IBM* System x* Beowulf Cluster consisting of 1,048 Intel* Xeon CPUs for parallel computations and several IBM System p* symmetric multiprocessing computer systems for memory-intensive computations. Typically, Beowulf Clusters are scalable-performance groups of usually identical PCs, running an open-source operating system (OS), such as Linux*. They're connected via a high-speed communication network and have common programming and inter-processor communication libraries that allow system resources to be shared for parallel computing.

"In the field of bioinformatics and computational biology, most open-source pro-

ILLUSTRATION BY CARDLE VERBYST

grams have been developed for use on Linux," says Edward B. Suh, Sc.D., CIO. "Biologists are comfortable and familiar with Linux. Open-source programs are readily available; and open-source code can help optimize your computational requirements rapidly on your hardware."

The center's machines are linked using a network comprised of a Gb Ethernet connecting 524 nodes (each node has two Intel Xeon CPUs). Additionally, the center has 128 of these nodes connected with Myrinet, a low-latency, high-speed interconnect from Myricom. "This hardware provides the ability for the nodes to communicate much faster with each other," says James Lowey, manager of the center. "This enables larger, more interdependent computational problems to be examined on this system."

For example, a lab may want to perform a test that would require another lab's research findings; perhaps statistically significant findings on diseased cells treated with a specific cancer drug, to see if that drug would be effective on another type of cancer with the same gene-expression pattern.

The ability to analyze these types of complex problems can be expedited by parallel computing with the use of MPICH, a popular message-passing interface (MPI) library, used in the Beowulf Cluster environment. MPICH was developed jointly by Argonne National Laboratory, under the direction of Bill Gropp and Rusty Lusk; and Mississippi State University, under Tony Skjellum and Nathan Doss. IBM also made major contributions to development of MPICH, under Hubertus Frank.

Patterns, Associations and Interactions

To fully appreciate the impact of these technological contributions on life-saving science, you have to understand the processes. "We look at cancer," says Suh. To be exact, the center's genomic research encompasses tumor classification, risk assessment and prognosis, drug development, drug response, therapy development and disease progression. "We may be looking at melanoma to see what genes play a role or are important to this cancer. The patterns we deal with are not only based on how often something appears, but also the interaction it causes in other genes' behaviors."

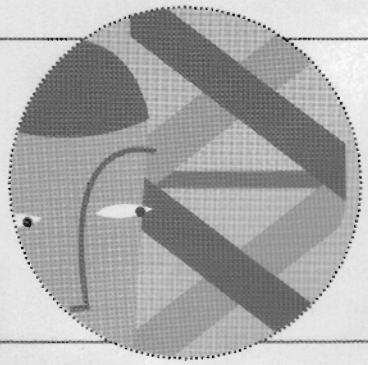
"How often" is key. To establish those frequencies, Suh and Lowey provide the scientists with machines that can run analysis on billions of gene combinations.

"The sheer volume of data that is generated by these techniques requires a huge amount of processing power, and without access to this computational power it would take many years to analyze the amount of data being generated," says Lowey.

The center, like other genetic-research labs, uses DNA microarrays, solid supports onto which the sequences from thousands of different genes are immobilized or attached to help them investigate genes' relationships to cancer. By arranging many short sequences of nucleotides, A, T, C and G, such that they complement the actual genes' sequences (for example, the complementary sequence to G-T-C-C-T-A will be C-A-G-G-A-T), scientists can measure the expressions of those genes over many thousands of DNA samples in a single experiment. They can then draw conclusions about the genes' behavior (e.g., comparing diseased cells with healthy ones).

To isolate what genes are involved in a cancerous cell growth, scientists use a special type of DNA microarray chip, to detect whether a particular gene or set of genes is being expressed more or less than others, under given circumstances. This is called microarray-expression analysis. This technique can be used to determine the correct treatment for a disease. By comparing the similar expression of a gene pair in a particular form of cancer, it's possible to compare the sample of diseased tissue from a patient and discover a match, thus making the diagnosis. This is why it's so important to be able to compare billions of combinations quickly. Also, by isolating the expression pattern, expression chips can be used to develop new drugs.

To examine just 600 genes, in order to see the relationship of four gene combinations to other genes, scientists may have to look at more than 5 billion gene combinations. "If each gene combination analysis takes the range of 0.05



"We are on the edge of an unprecedented quantum leap in medical science, dependent on the computational ability to perform complex data analyses and simulations as quickly as possible."

—Edward B. Suh, CIO, TGen

seconds, that's on the order of 10 years of computer analysis time," says Suh. "But if you have 100 CPUs working together, instead of 10 years, it would only take a month or two—or with 1,000 CPUs, less than a week."

"It's all about the compression of time," agrees Lowey. "If one problem takes four years to compute with one CPU, this cluster could enable it to run in just a couple of days." The computer runtimes will vary by application, of course, he adds. Some applications will have 90-percent parallel efficiency, while others may have only 50 percent. Parallel computing on a Beowulf Cluster enables the scientists to divide a compute-intensive large problem into small chunks that can be assigned to and run on many processors simultaneously. The level of parallel efficiency is based on the ratio of the additional increase in total computer runtime to the number of processors used.

The center is unique in that some of the microarrays it uses have up to 500,000 target spots on them. The data from the microarrays, which can build up very quickly, is stored in IBM General Parallel File Systems (GPFS). Application programs running on the Beowulf Cluster supercomputer draw from that data for analysis. To comb the data, researchers run it through a variety of computational algorithms, or chains of complex sets of orders.

"Most open-source scientific applications are developed in the Linux environment," says Suh. "Linux enables quicker development time."

One of the open-source applications that's frequently run on the center's Beowulf Cluster is the NAMD, a molecular modeling application developed by the University of Illinois at Urbana-Champlain.

The center may use the NAMD program to analyze proteins, which are important in cancer research because an abnormal protein signals that the gene that made it has mutated.

UP CLOSE

CUSTOMER: TGen

HEADQUARTERS: Phoenix

BUSINESS: Research institute

HARDWARE: System x Beowulf Cluster, System p hardware, General Parallel File Systems

SOFTWARE: Linux

CHALLENGE: Simplifying IT environment to help facilitate cancer research

SOLUTION: Using System p and System x hardware with Linux to tackle compute-intensive problems quickly

A protein is a three-dimensional structure encoded in the gene, which contains the DNA instructions on how that protein should be made up. Consequently, if this gene is mutated, the protein it produces will have an abnormal protein structure, produce abnormal hormones, etc.

To diagnose a protein, you need to unlock that structure. The NAMD program allows researchers to model the proteins by providing the tools to analyze the interaction between atoms, using the laws of thermodynamics. The researchers can observe how the protein moves by measuring the active and repulsive forces between atoms. Again, by comparing the behavior patterns of healthy proteins, the aberrant ones can be singled out.

If the protein structure is known, a drug can be developed that will attach itself to that shape. Without even touching the patient, researchers can computationally model the drug's interaction on a biomarker gene set—which is over expressed in cancer—to see if the new drug or an existing drug will reduce over expression and force the cancer into remission.

"What we would like to be able to do is to take a set of genes and say, this set has something to do with your cancer—to diagnose and prognosticate, and create therapeutics for each cancer," says Suh. "For certain types of cancers, we are already doing that now. So we have some success stories."

But Suh isn't content to stop at these successes; he's asking other questions, such as whether the same cancer drug that works on breast cancer will work on other cancers. This requires complicated analysis. But, as the saying goes, "The impossible will just take a little longer."

Limitless Work

"The amount of work that can be done is limitless," he says. "We are on the edge of an unprecedented quantum leap in medical science, dependent on the computational ability to perform complex data analyses and simulations as quickly as possible. The science, the equipment is here. The key is processing and manipulating large volumes of data at extreme computational speeds." The ASU-TGen High Performance Computing Center appears to be on the right path for this race against disease, through the center's open network and operating environment; and its acquisition of increasingly powerful processors.

Maureen Nevin *has covered information technology and finance since the 1980s. Maureen can be reached at MNevinDuffy@aol.com.*